HW3 - COMP 329: NLP

Matt Hyatt

March 17, 2022

Procedure

Preprocessing

I preprocessed the Movie dataset from HW1 first, by converting the entire corpus tp lowercase letters `.lower()` and then used regex `[^/w ]` to remove all punctuation from the text.

Model

I implemented a custom TFIDF vectorizer by calculating the product of TF and IDF for each word, document combination. I calculated TF by summing the frequency of each word in a document and dividing this by the number of total words in that document. IDF was the log of the number of documents N divided by the number of documents containing word W.

I used a multilayer perceptron to train my classifier. I decided that a slightly slower compute time was not worth reducing accuracy by reducing my feature dimensions, since the MLP will make good use of every available feature. For this reason, I decided to leave n and m unchanged, as 0 and 18920 respectively.

One problem I ran into was a divide by 0 error resulting from cases where the test set did not contain any words from the training data. I resolved the issue by adding 1 to the denominator of my TF calculation.

TF = frequency[word] len(document)+1

My best results were with the Adam optimizer and early stopping parameter, to keep the model from accidentally overfitting.

Results

My model received a validation accuracy of 77% and test accuracy of 77.8%. This is slightly better than my Naive Bayes classifier from Homework 2 (76% accuracy). When using scikit-learn's TfidfVectorizer, my model recieved a validation accuracy of 77.8% and test accuracy of 78.3%